Can Deep Learning Models Recognize Chilean Diet?

Bastián Muñoz, Ignacio Chirino and Eduardo Aguilar

***Abstract*—The emergence of deep learning models has made it possible to address several real-life problems and, in particular, those in which computer vision plays a key role. In this sense, the food recognition task from images is one of the beneficiaries of this machine learning method. Its importance lies in its usefulness to become aware of the food eaten and in this way help us to lead a healthy lifestyle. In recent years, food image recognition has gained great prominence in the literature, providing novel models and datasets to address it. However, public data generally correspond to American, Asian, and European foods, therefore the methods developed cannot be directly applied to the Chilean diet. In this article we will publish a new dataset to recognize the foods present in the Chilean diet. In addition, we will perform a comparison with public popular food datasets to analyze the similarity in the dishes of the proposed dataset with respect to the exciting ones in the literature. Moreover, we will establish a baseline using the state-of-the-arts Convolutional Neural Network architectures and the novel Swin Transformer approach.**

***Index Terms*—Food recognition, Food dataset, Chilean diet, Deep learning**

I. Introducción

A partir de la década pasada hemos estado experimentando un incremento sostenido en los resultados en múltiples problemas que pueden ser abordados mediante algoritmos de aprendizaje automático, especialmente con la aplicación de modelos de deep learning (DL). La reaparición y posterior rápida expansión de soluciones basadas en DL ha sido posible debido principalmente a tres factores habilitantes, los cuales proporcionaron el entorno perfecto para su desarrollo: 1) el aumento de las capacidades del hardware (procesador, memoria y tarjeta gráfica) a un costo asequible; 2) la enorme cantidad de datos digitales disponibles; y 3) los avances en herramientas informáticas para un eficaz desarrollo de métodos basados en redes neuronales.

La visión artificial es una de las disciplinas dentro de las ciencias de la computación que se ha visto fuertemente beneficiada con el surgimiento de DL, particularmente las redes neuronales convolucionales (CNN), evidenciando notorias mejoras en el desempeño en diversos dominios, tales como: reconocimiento de objetos [1], detección de objetos [2], generación y/o transformación de imágenes [3], por nombrar solo algunos. No hay duda alguna que este tipo de algoritmos han permitido proveer soluciones esperanzadoras a problemas de la vida real que eran resueltos parcialmente en condiciones controladas o simplemente eran impensados de abordar.

Bastián Muñoz, Ignacio Chirino and Eduardo Aguilar are with Department of Computer and Systems Engineering, Universidad Católica del Norte, Angamos 0610 Avenue, Antofagasta, Chile. (e-mail: bastian.munoz01@alumnos.ucn.cl and ignacio.chirino@alumnos.ucn.cl, eaguilar02@ucn.cl)

Específicamente, el análisis visual de alimentos es una de las problemáticas desafiante que se ha podido abordar con alentadores resultados mediante algoritmos de DL. Su importancia recae principalmente en concienciar a las personas sobre el daño que producimos a nuestra salud al mantener una dieta poco saludable. Tenga en cuenta que una mala alimentación puede producir diversos problemas de salud como la obesidad, diabetes, hipertensión, accidentes cerebro-vasculares, artrosis, algunos tipos de cáncer, entre otros [4].

Es importante destacar que la obesidad ha sido declarada como pandemia según la World Health Organization (WHO) [5]. Es una enfermedad que afecta a todo el mundo y a personas de todas las edades [6], disminuyendo la esperanza de vida y generando altos costos sociales y económicos para quienes la padecen. La obesidad tiene una estrecha relación con el desarrollo de otras enfermedades crónicas [4], [7]. Además, aumenta el riesgo de desarrollar síntomas graves en otras enfermedades, tales como hemos podido notar recientemente con el COVID-19 [8]. Particularmente en Chile, se ha observado una alta prevalencia y crecimiento de la obesidad en todas las etapas de la vida [9]. A nivel mundial, Chile se encuentra en el primer cuartil de los países con mayor porcentaje de obesidad, con más de un 25 % de la población adulta y más del 13 % de la población infantil [5].

Mayormente la obesidad es alcanzada debido a un desbalance entre las calorías que ingerimos con respecto a aquellas que quemamos, además de otros factores que tienen relación con mantener un estilo de vida sedentario. En este sentido, generalmente las personas obesas son tratadas para equilibrar su ingesta alimenticia. Para ello es necesario disponer de herramientas que permitan controlar y realizar el seguimiento de la dieta diaria. Los métodos tradicionales se basan en el auto-reporte, los cuales son propensos de presentar información ambigua, poco precisa e incompleta, sumado al tiempo requerido para completarlos [10]. Aquí recae la importancia del análisis visual de alimentos, facilitando la adquisición de la información y minimizando los problemas comentados.

Un sistema de control y monitoreo automático de la dieta a partir de imágenes de alimentos permite dotar a los pacientes de obesidad y especialistas de nutrición de una herramienta cómoda para mejorar la experiencia de seguimiento de una dieta saludable. Hay al menos tres desafíos que pueden ser abordados mediante algoritmos de DL en este tipo de sistema: a) Food Recognition, con el propósito de identificar uno o más comidas que estén presentes en el plato; b) Food Segmentation, para extraer la porción de cada alimento; c) Depth estimation, con el fin de ajustar la porción según la distancia en que la imagen fue tomada. Con estos datos, sumado con tablas

nutricionales estandarizadas de comidas, es posible ofrecer una adecuada cuantificación de la información nutricional relacionada a cada comida ingerida.

En el presente artículo nos enfocamos en el primer desafío (Food Recognition) y nos planteamos la siguiente pregunta: ¿Pueden los modelos de deep learning reconocer la dieta Chilena? Para responder a esta pregunta, es necesario explorar dos aspectos. El primero tiene relación con analizar si las imágenes públicas de alimentos actualmente disponibles en la literatura son lo suficientemente genéricas para ser aplicadas al reconocimiento de la comida chilena. El segundo corresponde a analizar la exactitud de diversos modelos basados en DL, que presentan los mejores desempeños en el estado del arte del reconocimiento de objetos, sobre el reconocimiento visual de alimentos pertenecientes a la dieta chilena.

Las contribuciones de nuestro artículo puede ser resumidas de la siguiente manera:

- Liberación de ChileanFood-64, un nuevo dataset con imágenes de comida típica Chilena etiquetado para abordar el problema del Food Recognition.
- Línea base en el reconocimiento de comida mediante modelos de DL basados en CNN o Transformer [11], sobre dos datasets públicos y sobre ChileanFood-64.
- Estudio comparativo sobre semejanza entre las comidas pertenecientes a ChileanFood-64 y las comidas disponibles en dos populares datasets públicos.

En la próxima sección se presenta una revisión del estado del arte en cuanto a los trabajos relacionados con los algoritmos de aprendizaje automático utilizados para el reconocimiento de comida así como también la disposición de nuevas bases de datos para este propósito. En la sección III, se describe en detalle el proceso de confección del conjunto de datos de comida chilena propuesto. En la sección IV, se presentan la configuración experimental, los resultados obtenidos por los modelos de DL seleccionados y el análisis de la similitud entre las clases contenidas en los conjuntos de datos populares con respecto a las pertenecientes a la comida chilena. Finalmente en la sección V, se exponen las conclusiones y el trabajo futuro.

II. Trabajos relacionados

En esta sección se describen los trabajos relacionados con nuestra investigación. En particular, los métodos desarrollados para el reconocimiento visual de la comida y los conjuntos de datos actualmente disponibles para este fin.

A. Food Recognition

Las primeras soluciones para el reconocimiento de comida se basan en características *hand-crafted* [12]–[14]. En estos trabajos, la extracción de características se realiza localmente (ej. SIFT) y/o globalmente (ej. basadas en el color), y se utilizan individualmente [12] o en combinación [13], [14]. Otros trabajos consideran estimar la relación entre distintos ingredientes de la imagen mediante métodos estadísticos para posteriormente realizar la clasificación [15], o el uso de información contextual como las coordenadas de la localización (GPS) e información particular del restaurante [14] para mejorar el desempeño. Estos métodos funcionan relativamente bien en situaciones donde el fondo de la imagen no tiene ruido y la comida esta bien posicionada. A diferencia de los métodos tradicionales que son difíciles de aplicar con imágenes de la vida cotidiana, surgen los métodos basados en DL. Tempranas investigaciones en DL muestran que las CNN superan en rendimiento a los métodos tradicionales por un amplio margen tanto para el reconocimiento de objetos en general [16] como para el reconocimiento de alimentos [17], [18].

Diversos trabajos han sido propuestos en los últimos años con el fin de seguir mejorando el desempeño. Estrategias de fusión temprana [19], es decir antes de la clasificación, y de fusión tardía [20], [21] se han evidenciado en la literatura. Concretamente, en [19] se propone el modelo WISeR que contempla la fusión temprana (concatenación) de las características extraídas mediante 2 redes, una de propósito general (WRN [22]) y la otra diseñada para extraer los rasgos verticales distintivos de las imágenes de comida. En cuanto a la fusión tardía, la clasificación final realizada a partir de los resultados individuales de cada arquitectura de CNN que conforman el *ensemble* ha sido realizada mediante *decision template scheme* [20] o mediante varios enfoques de votación [21]. Otra métodos que fusionan diversos tipos de características han sido propuestos en [23]–[25]. El método MSMVFA, propuesto en [23], contempla la extracción de características semánticas de alto nivel (platos reconocidos), características de atributos de nivel medio (ingredientes reconocidos) y características visuales profundas (características usadas para el reconocimiento del los platos) a partir de dos redes neuronales entrenadas con imágenes transformadas en múltiples escalas. Las características extraídas para cada escala son fusionadas según el tipo de características, luego el vector resultante es normalizado para posteriormente realizar la agregación y la clasificación de la representación obtenida. En cuanto al método PAR-Net, propuesto en [24], su estructura contempla el entrenamiento de tres redes en paralelo y la posterior concatenación de las características de cada una de ellas para realizar la clasificación. La salida de la primera red corresponde a la predicción de la imagen original. La salida de la segunda, a la predicción de la imagen con regiones discriminativas eliminadas. Y por último, la tercera a la predicción de las regiones discriminativas recortadas y re-escaladas al tamaño original. Para descubrir las regiones discriminantes es utilizada la técnica Adversarial Erasing [26]. Con respecto al método SGLANet [25], se propone una red que aprende conjuntamente características globales y locales complementarias las cuales son fusionadas para el reconocimiento de comida.

También ha sido investigado el aprendizaje de múltiples tareas con el propósito de mejorar el desempeño de la clasificación por medio del aprendizaje en conjunto de tareas relacionadas con la comida. Además de las anotaciones con respecto a la comida, se han utilizado los datos contendidos en las recetas (método de cocción, ingredientes) [27], [28], el estilo de la cocina [29] y la información calórica [30]. La clasificación jerárquica de la comida también ha sido abordada dentro del marco de aprendizaje de múltiples tareas [31], [32]. Este enfoque ha permitido obtener mejores errores, es decir, ocurridos en clases semánticamente similares [31], y también

mejorar los resultados de clasificación al integrar las relaciones visuales entre diferentes categorías de alimentos [32]. Por otro lado, la información jerárquica de la comida ha sido utilizada en el entrenamiento de un conjunto de redes especializadas, para posteriormente seleccionar aquella más acordé según el dato de entrada, para así realizar la clasificación final [33].

Más recientemente han sido propuestos métodos más avanzados para el reconocimiento de comida. Un enfoque basado en la destilación de conocimiento se propone en [34] con el fin de proveer de un algoritmo preciso y eficiente para ser utilizado en dispositivos móviles. Además, un método que aprende las relaciones entre las distintas clases mediante Graph Convolutional Network a partir de una codificación semántica de la etiqueta de la clase y la fusión del aprendizaje basado en múltiples (*many-shot*) o pocos datos (*few-shot*) se propone en [35]. Del mismo modo, métodos de DL de propósito general publicados últimamente [1], [36] han demostrado ser beneficiosos para el reconocimiento de comida.

De la revisión de la literatura, podemos ver que los conjuntos de datos generalmente utilizadas para hacer el estudio comparativo en el reconocimiento de comida son UECFood-256 [17], [21] y Food-101 [18], [36], y además las arquitecturas basadas en CNN ResNet [19], [20] y DenseNet [17], [23] son frecuentemente utilizadas mostrando un destacado desempeño.

B. Datasets

Pittsburgh Fast-Food Image Dataset (PFID) [12] es el primer conjunto de datos público de imágenes de comida, compuesto principalmente de comida rápida. El pobre desempeño obtenido por los métodos tradicionales de visión artificial que se usaban popularmente en ese momento demostró la complejidad del problema del reconocimiento de alimentos.

Luego de la exitosa aparición de los modelos de DL y el fácil acceso a imágenes de comidas, fueron surgiendo conjuntos de datos de comida más desafiantes y con una mayor cantidad de platos. Este es el caso de Food-101 [37] publicado en 2014 conteniendo 101 categorías de comida internacional, teniendo 1.000 imágenes por clase y un total de 101.000 imágenes. En el mismo año, se publica UECFood-100 [38], un conjunto de datos de comida Japonesa que inicialmente contenía 9.000 imágenes de 100 categorías distintas, y que fue posteriormente expandido logrando un total de 256 comidas incluyendo algunas comidas internacionales, publicado bajo el nombre de UECFood-256 [39]. Ambos datasets son frecuentemente utilizados para los estudios comparativos dentro del marco del reconocimiento de comida.

Otras conjuntos de datos públicos de imágenes de comida proveen de información adicional al nombre de la comida [27], [29], contienen imágenes de comidas y bebidas [40], o imágenes de frutas y verduras [41]. Específicamente, VIREOFood-172 [27], que consiste de 110.241 imágenes divididas en 172 distintas categorías representativas de la comida China, dispone de anotaciones del plato y también de los ingredientes que lo conforman desde un total de 353 ingredientes disponibles. MAFood-121 [29], que contiene un total de 21.175 imágenes distribuidos en platos populares pertenecientes a 11 estilos de cocina, provee de anotaciones para tres tareas relacionadas con la comida: el nombre del plato, el estilo de la cocina y el grupo de comida. En 2017 fue publicado un conjunto de imágenes de comida [40] con 520 diferentes categorías contenidas en un total de 225.953 imágenes en las que se incluyen comidas y bebidas. Finalmente, en el mismo año se publico el conjunto de datos VegFru que se caracteriza por contener vegetales y frutas con anotaciones que incluyen 2 niveles de jerarquía y un total de 160.000 imágenes.

También se han publicado conjuntos de imágenes de comida perteneciente a algún estilo de comida en particular [32], [42], [43]. El conjunto de imágenes de comida THFOOD-50 [42], fue publicado el 2017 y contiene 50 comidas populares tailandesas con una cantidad de imágenes aproximadamente entre 200 y 700 para cada categoría. Por otro lado, TürkSofrası-15 [43] es un conjunto de imágenes de comida turca que contiene 15 comidas típicas con aproximadamente 500 imágenes para cada categoría. Por último, en el año 2020 se publica VIPER-FoodNet (VFN) [32], un dataset que contiene 82 categorías de comida Americana a lo largo de 14.991 imágenes.

Recientemente, se ha publicado el conjunto de imágenes de comida ISIA Food-500 [25], que posee 500 categorías distintas y aproximadamente 399.726 imágenes en total. Sobrepasando así a a los datasets mas populares tanto por la cantidad de categorías como en cantidad de imágenes. Además, se han publicado conjuntos de imágenes con anotaciones a nivel de píxel, con el propósito de entrenar modelos de segmentación de comida [44]–[46]. Particularmente, FOOD50SEG [44] es uno de los primeros conjuntos de imágenes que provee anotaciones a nivel de píxeles en una gran cantidad de datos, en concreto, sobre un total de 120.000 imágenes distribuidos en 50 categorías distintas. Otros conjuntos de datos liberados para la segmentación son UECFoodPix [45] y UECFoodPix Complete [46], con un total de 10.000 imágenes distribuidas en 102 clases de comida. Las anotaciones se obtienen de manera no supervisada a partir de las anotaciones de las regiones donde están presentes las imágenes (*bounding boxes*) [45] o etiquetadas manualmente a nivel de píxel [46].

Como se observa de la revisión de la literatura (ver resumen en Tabla I), no se encuentra evidencia de conjunto de imágenes de comidas que representen la cocina Chilena.

III. CHILEANFOOD-64 DATASET

En esta sección se describe en detalle el procedimiento realizado para la confección del conjunto de datos ChileanFood-64 compuesto de platos de comida típica de la dieta Chilena.

A. Recopilación de los Datos

El primer paso consintió en la recopilación de las categorías candidatas pertenecientes a la comida Chilena y la recuperación de imágenes para cada una de ellas. Teniendo en cuenta la experiencia personal del equipo sumado con una búsqueda exhaustiva en Internet, fuimos capaces de obtener un total de 74 categorías entre diversas comidas, postres y bebidas típicas de la comida Chilena. Lo siguiente fue la recolección de imágenes para cada una de las categorías elegidas. Para ello, al igual que la mayoría de los conjuntos de datos publicados [29], [32], [39], [40], se realizaron búsquedas en Internet (Google

TABLA I
LISTA DE CONJUNTO DE DATOS UTILIZADOS EN LA LITERATURA.

Nombre	Año	Imágenes	Clases	Anotación	Cocina
PFID	2009	4.545	101	Label	USA
UECFood-100	2012	9.060	100	BBox	Japones
Food-101	2014	101.000	101	Label	Universal
UECFood-256	2014	31.397	256	Label/BBox	Japones y otros
VIREO	2016	110.241	172	Label/Receta	China
NutriNet	2017	225.953	520	Label	Europa Central
VegFru	2017	160.000	292	Label	Universal
THFOOD-50	2017	15.770	50	Label	Tailandés
TürkSofrası-15	2017	7.500	15	Label	Turca
MAFood-121	2019	21.175	121	Label	Universal
UEC-FoodPix	2019	10.000	102	Region/Label	Japones
FOOD50SEG	2020	120.000	50	Region/Label	Universal
ISIAFood-500	2020	399.726	500	Label	Universal
VIPER	2020	14.991	82	BBox	USA
UEC-FoodPix Comp.	2020	10.000	102	Region/Label	Japones

Images, Bing e Instagram) mediante algoritmos optimizados (*web scraping*), usando extensiones de los navegadores web y también manualmente. En la búsqueda se utilizo literalmente el nombre del plato, traducciones y pequeños ajustes (eliminación o incorporación de palabras) como palabras clave. Dando como resultado en la mayoría de las clases más de 600 imágenes. Aquellas clases en que se recuperaron menos de 600 imágenes, fueron descartadas inmediatamente. Como resultado de esta etapa quedaron un total de 64 categorías de comida de variados tipos: bebidas, carnes, ensaladas, fritos, mariscos, masas, mezcla, postre, sopa y verduras cocidas.

B. Limpieza y Etiquetado de Datos

Al realizar un proceso automático de recolección de imágenes es muy probable la obtención de datos corruptos o incorrectos para una categoría en particular. Por lo tanto, resulta obligatorio hacer una revisión de los datos para limpiar aquellas imágenes no deseadas y con ello asegurar la calidad del conjunto de datos propuesto.

La primera tarea realizada en esta etapa fue la eliminación de imágenes corruptas. Esto se hizo automáticamente usando la librería PIL compatible con Python3 mediante la función de nombre *verify*, la cual genera una excepción cuando los datos tienen algún problema. Todos aquellos datos que generaron la excepción se consideraron corruptos. Posteriormente, la segunda tarea, corresponde la eliminación de las imágenes descargadas de manera duplicadas. Para ello, en primer lugar se realiza un reajuste de tamaño a cada una de las imágenes obteniendo como resultado imágenes de dimensión 30×30. Entonces, con las imágenes re-dimensionadas, se prosigue con la normalización de cada una de ellas mediante la estrategia de normalización *min-max*. Luego, mediante la distancia de Manhattan se estima la similitud entre pares de imágenes. Tenga en cuenta que un par de imágenes exactamente iguales les corresponde una similitud $L_1 = 0$. Sin embargo, un par de imágenes iguales pueden tener pequeñas diferencias producto a factores aleatorios, por lo que decidimos flexibilizar

Fig. 1. Ejemplo de comida contenida en ChileanFood-64.

el criterio y asumir que dos imágenes son semejantes cuando $L_1 < \delta$, con el umbral $\delta = 0{,}02$. Con esto, mediante un proceso iterativo se comparan todas las imágenes y se elimina aquellas que se encuentren bajo el umbral de similitud. La tercera etapa de limpieza consistió en la eliminación de imágenes que no representan a un plato de comida. Con el fin de agilizar el proceso de limpieza, se implemento un clasificador binario para poder distinguir entre imágenes que eran catalogadas como *comida* o *no comida*. El clasificador se entreno utilizando como conjunto de datos el dataset Food-5K, logrando un exactitud aproximada del 99 %. Sorprendentemente, al probar este clasificador en nuestro conjunto de datos observamos que se descartaba un numero alto de imágenes de comida, a pesar del buen desempeño mostrado sobre el conjunto de datos usado para el entrenamiento. Por esta razón decidimos continuar con el proceso de filtrado manual en donde se revisaron detenidamente todas las imágenes para asegurar su pertenencia para cada clase. En este proceso de filtrado manual se eliminaron aquellas imágenes que no corresponden a una comida y se conservaron aquellas que satisfacen estos criterios: a) Deben tener claridad, b) Deben enfocarse en la comida de la clase etiquetada y c) No deben tener ningún tipo de distorsión o modificación.

Una vez acabada la limpieza del conjunto de datos, nuevamente aplicamos el clasificador de comida/no comida entrenado previamente. Se esperaría una baja predicción para la clase no comida, sin embargo sucedió todo lo contrario. Esto sugiere que hay una notoria diferencia entre los platos típicos de comida Chilena con respecto a los platos utilizados para generar la base de datos Food-5K.

Por último, con el conjunto de datos limpio, se formatea para los nombres de las carpetas e imágenes y se generan las anotaciones para cada imagen de acuerdo al nombre del plato que representan. Las anotaciones son almacenadas en archivos *CSV* con dos columnas, una para indicar la ruta relativa del fichero con respecto al directorio en donde se encuentran las imágenes y la otra para indicar la etiqueta respectiva.

C. Datos Resultantes

Como resultado del proceso de adquisición de datos se completo la elaboración de un conjunto de datos de comida Chilena, con un total 11.504 imágenes distribuidas a lo largo de 64 clases distintivas de esta cocina (ver Fig. 1).

En la Fig. 2 se puede observar la cantidad de imágenes resultantes para cada categoría. Se interpreta que a pesar de haber recopilado una gran cantidad de datos para cada clase, luego del proceso de limpieza, el conjunto de datos quedó

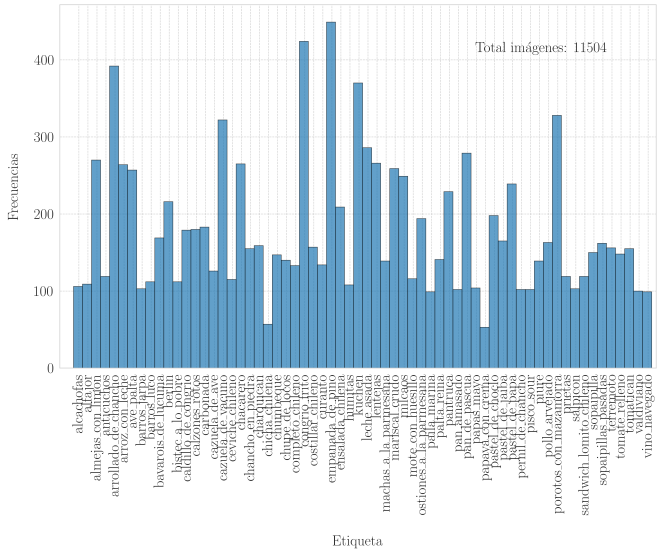

Fig. 2. Distribución de las imágenes para cada clases de ChileanFood-64.

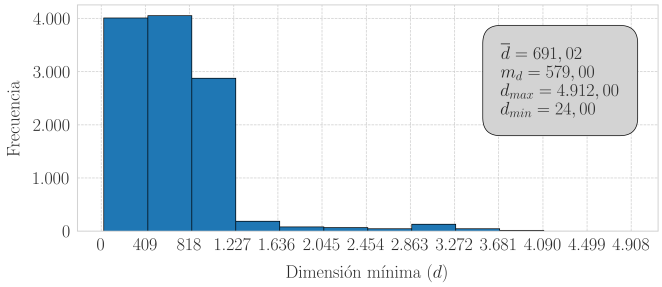

Fig. 3. Distribución de las resoluciones de las imágenes de acuerdo a la dimensión mínima.

des-balanceado. Sin embargo, es interesante de resaltar que la mayoría de las clases poseen a lo menos 100 imágenes.

Por otro lado, en la Fig. 3 se puede observar un histograma que refleja la frecuencia de imágenes contenidas en ChileanFood-64 pertenecientes a distintos rangos de resoluciones y también algunas métricas estadísticas tales como la media aritmética ($\overline{d}$), la mediana (m_d), el valor máximo (d_{max}) y el valor mínimo (d_{min}). Para la construcción de este histograma se considera la dimensión mínima d de una imagen sin importar si corresponde a la dimensión vertical (ancho) u horizontal (alto) de la misma. De los datos podemos interpretar que existen imágenes de muy baja resolución (lado mínimo 24 píxeles) o de muy alta resolución (lado mínimo 4.912 píxeles), predominan las imágenes en que d es a lo mucho 1.227 píxeles y la mitad de las imágenes tiene como máximo un d igual a 579 píxeles. La amplia variedad de resoluciones sugiere un desafió extra para el aprendizaje de los modelos DL.

Finalmente, en cuanto a la distribución de los datos para el entrenamiento, se proporcionan tres conjuntos (*train*, *val* y *test*) obtenidos a partir de un muestreo aleatorio estratificado sobre el total de los datos. Las imágenes se dividieron inicialmente en un 85 % para entrenamiento (*train*) y en un 15 % para pruebas (*test*). Luego, nuevamente se realiza el muestreo estratificado sobre el conjunto de *train*, separando este conjunto en el conjunto definitivo de entrenamiento y en el de validación (*val*). En resumen, los datos se dividen en un 72.25 % para *train*, un 12.75 % para *val* y un 15 % para *test*.

IV. Experimentos

En esta sección se presenta la configuración experimental utilizada para el entrenamiento de diversos algoritmos de DL basados en arquitecturas CNN y Transformer. Luego, se presentan los resultados obtenidos por los modelos entrenados sobre tres bases de datos de comida. Finalmente, se describen los resultados obtenidos a partir del análisis de la semejanza entre los conjuntos de datos públicos y el propuesto.

A. Configuración Experimental

Para la experimentación consideramos el entrenamiento de 4 métodos de DL que han entregado excelentes resultados en el reconocimiento de objetos en los últimos 5 años, 3 de ellos corresponden a arquitecturas basadas en CNN (ResNet-50 [47], DenseNet-169 [48] y EfficientNetB0 [1]) y un cuarto basado en Transformer (Swin-T [49]). Las versiones de estos métodos se seleccionaron considerando la resolución de la imagen de entrada, específicamente, en nuestro experimentos utilizamos una resolución de entrada de 224×224 píxeles.

Con el fin de garantizar que los resultados sean directamente comparables, se realiza exactamente el mismo proceso de entrenamiento para todos los métodos. Todas los métodos son pre-entrenados en ImageNet [16] y luego se entrenan durante 40 épocas, con un *batch-size* de 16, *learning rate* de $2e^{-4}$ que se divide en 5 cada 8 épocas y una paciencia de 8 épocas. En cuanto a los datos, las imágenes inicialmente se redimensionan a un tamaño de 256×256. Posteriormente, se aplican tradicionales técnicas de aumento de datos (ej. rotación, desplazamiento, volteo horizontal, etc) y recortes aleatorios de tamaño 224×224. Por último todos los píxeles se escalan por un factor de $1/255$, quedando normalizados en un rango de $[0 - 1]$. El entrenamiento se realiza minimizando la perdida de entropía cruzada usando Adam como optimizador en el Framework Keras con Tensorflow como Backend.

En lo que concierne a la evaluación del desempeño de los métodos de DL, se seleccionan 2 conjuntos de datos populares para el reconocimiento de alimentos (Food-101 y UECFood-256) y el propuesto (ChileanFood-64), y se usa como instrumento de evaluación la métrica tradicional elegida para la evaluación de modelos de reconocimiento de objetos, $Accuracy$ (Acc). En concreto, se calcula Top1 y Top5 Acc teniendo en cuenta los resultados de la predicción de las imágenes redimensionadas en 256×256 y recortadas en 224×224 píxeles en el centro de la imagen.

B. Resultados en el Reconocimiento de Comida

Los resultados obtenidos en el reconocimiento de alimentos por los 4 métodos de DL seleccionados sobre las bases de datos Food-101, UECFood-256 y ChileanFood-64 se muestran en la Tabla II. Se puede apreciar que DenseNet-169 y Swin-T son los métodos que proveen el mejor desempeño. En cuanto a EfficientNetB0, el rendimiento es cercano a los mejores métodos y además aprovecha mejor los recursos computacionales

en cuanto a la memoria requerida para su almacenamiento y el tiempo de predicción (50 % más rápido que DenseNet-169 y 35 % más rápido que Swin-T). Con respecto a ResNet-50, se observa en la experimentación que predice en un tiempo equivalente al de EfficientNetB0, sin embargo los resultado en términos de *Acc* son muy inferiores. En resumen, si los recursos computaciones y el tiempo no son una limitante, y se considera imágenes con resolución de 224×224, Swin-T es el método que muestra el mejor desempeño en el reconocimiento. En cuanto al nivel de dificultad de ChileanFood-64, se observa que es equiparable con conjuntos de datos populares. Aunque hay menos clases que en Food-101, el hecho de contener datos desbalanceados y menos de 1.000 muestras por clase (como en UECFood-256), aumenta la complejidad del aprendizaje.

TABLA II
RESULTADOS SOBRE FOOD-101, UECFOOD-256 Y CHILEANFOOD-64 EN TÉRMINOS DE TOP1 Y TOP5 *Acc*.

Modelos	Food-101		UECFood-256		ChileanFood-64	
ResNet-50	82,37 %	95,96 %	63,97 %	89,24 %	75,43 %	93,05 %
DenseNet-169	85,20 %	96,86 %	69,50 %	92,25 %	79,90 %	95,08 %
EfficientNetB0	84,80 %	96,90 %	68,94 %	91,94 %	78,79 %	95,48 %
Swin-T	86,43 %	97,25 %	69,21 %	92,04 %	80,59 %	95,60 %

C. Comparativa entre Datos Públicos y Datos Propuestos

Se analiza la semejanza entre las clases contenidas entre los conjuntos de datos populares (UECFood-256 y Food-101) con respecto a las contenidas en ChileanFood-64 con la intensión de determinar la existencia o ausencia de platos visualmente similares. Tenga en cuenta que de existir una alta correspondencia, no sería necesario publicar un nuevo conjunto de datos. Esto último debido a que bastaría simplemente con ajustar el nombre de la categoría de los conjuntos públicos para asociarlo a los platos contenidos en ChileanFood-64.

Para este análisis se realizan dos pasos previos. En primer lugar, se seleccionan los modelos que mejor desempeño lograron sobre las conjuntos de datos UECFood-256 y Food-101. Luego, utilizando los modelos seleccionados, se generan predicciones sobre las imágenes pertenecientes a ChileanFood-64 y se contabiliza la frecuencia de los platos predichos para cada categoría (ver primera columna en Fig. 4). Con ello, se inicia el análisis considerando 4 posibles casos que se pueden presentar. El primer caso corresponde a una correspondencia mayoritaria y clases visualmente similares, es decir, cuando se presenta una alta frecuencia en la predicción para un determinado plato y además este posee rasgos visualmente similares con respecto a la categoría objetiva (ver primera fila de la Fig. 4). El segundo caso corresponde a una correspondencia minoritaria y clases con baja similitud visual, es decir, no existe una predicción predominante y los platos son visualmente distintos (ver segunda fila de la Fig. 4). El tercer caso corresponde a una correspondencia mayoritaria y clases con baja similitud visual, es decir, se presenta una alta frecuencia en la predicción para un determinado plato, pero las imágenes son distintas visual y semánticamente (ver tercera fila de la Fig. 4). Por último, el cuarto caso, corresponde a una correspondencia minoritaria y clases con alta similitud visual, es decir, no se obtienen en las predicciones un plato predominante (no hay una clara correspondencia) a pesar de existir clases visualmente similares entre los conjuntos de datos (ver cuarta fila de la Fig. 4).

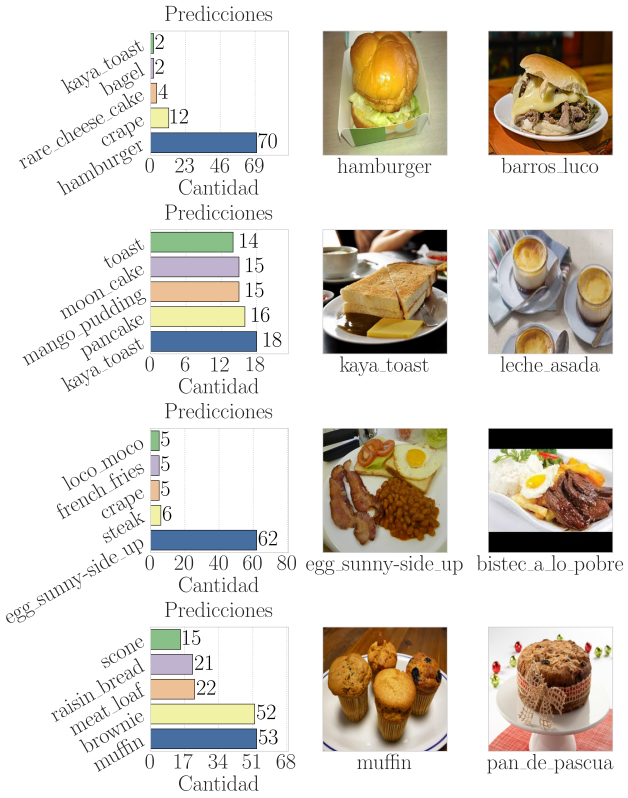

Fig. 4. Ilustración de los 4 casos evidenciados al comparar las clases pertenecientes a UECFood-256 con las propuestas. De izquierda a derecha se muestran: Histograma con la frecuencia de las clases predichas, predicción mayoritaria según las clases de UECFood-256 y clase objetiva de ChileanFood-64.

Todos los casos fueron ilustrados con los datos obtenidos sobre el conjunto de datos UECFood-256, sin embargo, se observan los mismos fenómenos sobre Food-101. En términos cuantitativos, si usáramos la clase mayoritaria para realizar las predicciones sobre nuestros datos, el resultado que obtendríamos seria de un 25.82 % y 18.13 % con los modelos entrenados sobre Food-101 y UECFood-256, respectivamente. Tenga en cuenta que en algunos casos la correspondencia no es única, por ejemplo, se observa que la clase hamburger resulta mayoritaria para barros luco, chacarero, sandwich lomito chileno. Lo que sugiere un resultado peor al indicado.

V. CONCLUSIONES

En este articulo se aborda el análisis de la pertinencia de los datos públicos actuales, así como, los métodos de DL para poder reconocer los alimentos consumidos en Chile. Por este motivo, se presenta ChileanFood-64, un nuevo conjunto de imágenes compuesto de comidas y bebidas típicas de Chile. En la experimentación se pudo observar que existen diversos casos donde se logra encontrar semejanzas bastante evidentes entre algunos platos de la comida chilena y platos existentes

en los conjuntos de datos públicos. Sin embargo, en la mayoría de casos esto no es posible. En cuanto a los métodos de DL, se establece una línea base usando algoritmos basados en CNN o Tranformer, en la que pudimos observar un desempeño de un 80 % sobre ChileanFood-64. Este desempeño resulta alentador de cara a un sistema automático para controlar la ingesta alimenticia. A futuro se pretende extender el conjunto de datos para tratar con la detección y segmentación de la comida.

REFERENCIAS

[1] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. on Mach. Learn.*, pp. 6105–6114, PMLR, 2019.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[3] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[4] I. Kyrou, H. S. Randeva, C. Tsigos, G. Kaltsas, and M. O. Weickert, "Clinical problems caused by obesity," *Endotext [Internet]*, 2018.

[5] "World obesity, obesity: missing the 2025 global targets – trends, costs and country reports," 2020.

[6] C. M. Apovian, "The obesity epidemic—understanding the disease and the treatment," 2016.

[7] R. Hakkak and A. Bell, "Obesity and the link to chronic disease development," *J Obes Chronic Dis*, vol. 1, no. 1, pp. 1–3, 2016.

[8] B. M. Kuehn, "More severe obesity leads to more severe covid-19 in study," *JAMA*, vol. 325, no. 16, pp. 1603–1603, 2021.

[9] F. Petermann-Rocha, M. A. Martínez-Sanguinetti, M. Villagrán, N. Ulloa, G. Nazar, C. Troncoso-Pantoja, A. Garrido-Méndez, L. Mardones, F. Lanuza, A. M. Leiva, *et al.*, "Desde una mirada global al contexto chileno:¿ qué factores han repercutido en el desarrollo de obesidad en chile?(parte 1)," *Revista chilena de nutrición*, vol. 47, no. 2, pp. 299–306, 2020.

[10] F. E. Thompson and A. F. Subar, "Dietary assessment methodology," *Nutrition in the Prevention and Treatment of Disease*, pp. 5–48, 2017.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[12] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," in *2009 16th IEEE ICIP*, pp. 289–292, IEEE, 2009.

[13] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *2011 18th IEEE ICIP*, pp. 1789–1792, IEEE, 2011.

[14] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *2015 IEEE WACV*, pp. 580–587, IEEE, 2015.

[15] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *2010 IEEE Computer Society Conf. CVPR*, pp. 2249–2256, IEEE, 2010.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[17] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proc. 2nd Int. Workshop MADiMa*, pp. 41–49, 2016.

[18] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *2015 IEEE ICMEW*, pp. 1–6, IEEE, 2015.

[19] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *2018 IEEE WACV*, pp. 567–576, IEEE, 2018.

[20] E. Aguilar, M. Bolaños, and P. Radeva, "Food recognition using fusion of classifiers based on cnns," in *ICIAP*, pp. 213–224, Springer, 2017.

[21] E. Tasci, "Voting combinations-based ensemble of fine-tuned convolutional neural networks for food image recognition," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30397–30418, 2020.

[22] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[23] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 265–276, 2019.

[24] J. Qiu, F. P. W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," 2019.

[25] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and S. Jiang, "Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proc. 28th ACM Int. Conf. on Multimedia*, pp. 393–401, 2020.

[26] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. CVPR*, pp. 1568–1576, 2017.

[27] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. 24th ACM Int. Conf. on Multimedia*, pp. 32–41, 2016.

[28] J.-j. Chen, C.-W. Ngo, and T.-S. Chua, "Cross-modal recipe retrieval with rich food attributes," in *Proc. 25th ACM Int. Conf. on Multimedia*, pp. 1771–1779, 2017.

[29] E. Aguilar, M. Bolaños, and P. Radeva, "Regularized uncertainty-based multi-task learning model for food analysis," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 360–370, 2019.

[30] T. Ege and K. Yanai, "Simultaneous estimation of food categories and calories with multi-task cnn," in *2017 15th IAPR Int. Conf. MVA*, pp. 198–201, IEEE, 2017.

[31] H. Wu, M. Merler, R. Uceda-Sosa, and J. R. Smith, "Learning to make better mistakes: Semantics-aware visual food recognition," in *Proc. 24th ACM Int. Conf. on Multimedia*, pp. 172–176, 2016.

[32] R. Mao, J. He, Z. Shao, S. K. Yarlagadda, and F. Zhu, "Visual aware hierarchy based food recognition," in *ICPR Workshops (5)*, pp. 571–598, 2020.

[33] E. Aguilar and P. Radeva, "Uncertainty-aware integration of local and flat classifiers for food recognition," *Pattern Recognition Letters*, vol. 136, pp. 237–243, 2020.

[34] H. Zhao, K.-H. Yap, A. C. Kot, and L. Duan, "Jdnet: A joint-learning distilled network for mobile visual food recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 665–675, 2020.

[35] H. Zhao, K.-H. Yap, and A. C. Kot, "Fusion learning using semantics and graph convolutional network for visual food recognition," in *Proc. IEEE/CVF WACV*, pp. 1711–1720, 2021.

[36] A. Bera, Z. Wharton, Y. Liu, N. Bessis, and A. Behera, "Attend and guide (ag-net): A keypoints-driven attention-based deep network for image recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3691–3704, 2021.

[37] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, pp. 446–461, Springer, 2014.

[38] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *2012 IEEE ICME*, pp. 25–30, IEEE, 2012.

[39] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *ECCV*, pp. 3–17, Springer, 2014.

[40] S. Mezgec and B. Koroušić Seljak, "Nutrinet: a deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, p. 657, 2017.

[41] S. Hou, Y. Feng, and Z. Wang, "Vegfru: A domain-specific dataset for fine-grained visual categorization," in *Proc. IEEE ICCV*, pp. 541–549, 2017.

[42] C. Termritthikun, P. Muneesawang, and S. Kanprachar, "Nu-innet: Thai food image recognition using convolutional neural networks on smartphone," *JTEC*, vol. 9, no. 2-6, pp. 63–67, 2017.

[43] C. Güngör, F. Baltacı, A. Erdem, and E. Erdem, "Turkish cuisine: A benchmark dataset with turkish meals for food recognition," in *2017 25th SIU Conf.*, pp. 1–4, IEEE, 2017.

[44] S. Aslan, G. Ciocca, D. Mazzini, and R. Schettini, "Benchmarking algorithms for food localization and semantic segmentation," *Int. J. of Mach. Learn. and Cybernetics*, vol. 11, no. 12, pp. 2827–2847, 2020.

[45] T. Ege, W. Shimoda, and K. Yanai, "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proc. 5th Int. Workshop MADiMa*, pp. 82–87, 2019.

[46] K. Okamoto and K. Yanai, "Uec-foodpix complete: A large-scale food image segmentation dataset.," in *ICPR Workshops (5)*, pp. 647–659, 2020.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR*, pp. 770–778, 2016.

[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. CVPR*, pp. 4700–4708, 2017.
[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

Bastián Muñoz Ordenes is a Civil Engineer in Computing and Informatics from the Universidad Católica del Norte, Antofagasta, Chile. His main interest in research area is Computer Vision using Deep Learning techniques.

Ignacio Chirino Farías is an engineering student and is currently opting for a university degree in Civil Engineering in Computing and Informatics at Universidad Católica del Norte, Antofagasta, Chile. His main interest in research area is Computer Vision using Deep Learning techniques.

Eduardo Aguilar Torres is a Doctor in Mathematics and Computer Science from the Universitat de Barcelona under the supervision of Dr. Petia Radeva. He is a Civil Engineer in Computing and Informatics and a Master's in Computer Engineering from the Universidad Católica del Norte. He is currently an academic in the Department of Computer and Systems Engineering at the Universidad Católica del Norte. His main interest is in the research and application of Deep Learning algorithms for visual food analysis. He aims to contribute to improving the quality of life of people through the generation of technological solutions based on Machine Learning and Computer Vision.